

# **Unsupervised Data Analytics in Mining Big Building Operational Data for Energy Efficiency Enhancement: A Review**

Cheng Fan<sup>a,\*</sup>, Fu Xiao<sup>b</sup>, Zhengdao Li<sup>a</sup>, Jiayuan Wang<sup>a</sup>

<sup>a</sup>Department of Construction Management and Real Estate, Shenzhen University,

Shenzhen, China

<sup>b</sup>Department of Building Services Engineering, The Hong Kong Polytechnic

University, Kowloon, Hong Kong, China

\*E-mail: [fancheng@szu.edu.cn](mailto:fancheng@szu.edu.cn); Tel.: (86) 26916426

## **Abstract**

Building operations account for the largest proportion of energy use throughout the building life cycle. The energy saving potential is considerable taking into account the existence of a wide variety of building operation deficiencies. The advancement in information technologies has made modern buildings to be not only energy-intensive, but also information-intensive. Massive amounts of building operational data, which are in essence the reflection of actual building operating conditions, are available for knowledge discovery. It is very promising to extract potentially useful insights from big building operational data, based on which actionable measures for energy efficiency enhancement are devised.

Data mining is an advanced technology for analyzing big data. It consists of two main types of data analytics, i.e., supervised and unsupervised analytics. Despite of the

power of supervised analytics in predictive modeling, unsupervised analytics are more practical and promising in discovering novel knowledge given limited prior knowledge. This paper provides a comprehensive review on the current utilization of unsupervised data analytics in mining massive building operational data. The commonly used unsupervised analytics are summarized according to their knowledge representations and applications. The challenges and opportunities are elaborated as guidance for future research in this multi-disciplinary field.

**Keywords:** Unsupervised data mining; Big data; Building operational performance; Building energy management; Building energy efficiency

## **1. Introduction**

The building industry has a significant impact on global sustainability. As indicated by the International Energy Agency (IEA), buildings have become the largest energy consumer in the world, accounting for over one-third of final energy consumption and an equally essential contributor of carbon dioxide emissions [1]. Building operations contribute to 80-90% of the total energy use throughout the building life cycle [2]. The energy saving potential is considerable taking into account the existence of a wide variety of building operation deficiencies. Various technologies have been developed to improve building operational performance. One prominent example is the Building Automation System (BAS), which is a network of hardware devices (e.g., servers, workstations, digital controllers and sensors) and software (e.g., building energy management programs and network communication protocols). BAS enables

building operations to be more intelligent by providing real-time monitoring and controls over different building services systems. A recent report showed that the energy saving potential from the adoption of advanced BAS might reach 22% by 2028 for the European building sector [3]. A typical BAS has the ability to record a large number of measurements and control signals at short time intervals (e.g., 30-second or 1-minute). As a result, massive amounts of building operational data, which are in essence the reflection of actual operating conditions, are available for data analysis. The knowledge hidden can be very helpful for a diversity of tasks in building energy management, such as predictive modeling, fault detection and diagnosis, and control optimization.

Some studies have been performed to develop advanced data analysis methodologies for mining massive building operational data. It is realized that conventional analytics, such as statistical and physical principle-based methods, are neither efficient nor effective in handling massive data sets. As a promising solution, data mining (DM) has drawn increasing attention due to its excellence in knowledge discovery from big data. DM is a multi-disciplinary subject, integrating techniques from statistics, machine learning, artificial intelligence, high performance computing and etc. There are two general types of DM, i.e., supervised and unsupervised analytics. While supervised analytics are powerful in modeling complicated relationships, unsupervised analytics are more practical and promising to discover novel knowledge given limited prior knowledge. Unsupervised analytics focus on exploring the intrinsic structures, correlations, associations and patterns in data and therefore, have

the ability to discover potentially useful yet previously unknown knowledge. More importantly, the success of implementing unsupervised analytics is less dependent on domain expertise and not subject to the availability of high-quality labeled training data. It is therefore reasonable to claim that unsupervised analytics will play an essential role in the upcoming big data era of the building industry.

Compared to supervised analytics, unsupervised analytics are less known and used in the building industry. This paper presents a comprehensive review on the current utilization of advanced unsupervised analytics in mining building operational data. It aims to provide a clear picture of the status quo of unsupervised analytics for building energy management, based on which focused research can be performed in the future. The paper is organized as follows. Section 2 serves as an introduction of unsupervised analytics. Section 3 discusses the typical types of data and knowledge embedded at the building operation stage. Section 4 reviews the research and applications in mining building operational data using unsupervised analytics. Section 5 presents the challenges and possible directions for future research. Conclusion is drawn in the last section.

## **2. Basics of unsupervised data analytics**

### **2.1 Supervised and unsupervised data analytics**

DM analytics can be generally classified into two categories, i.e., supervised and unsupervised analytics [4]. Supervised analytics, such as boosting and bootstrap aggregating, are powerful for predictive modeling. The knowledge representations of

supervised analytics are regression or classification models, which describe the quantitative or qualitative relationships between input and output variables. The success of supervised analytics is dependent on two factors, i.e., domain expertise and training data. Domain expertise is crucial for developing functional models. It is especially important for specifying model architecture, selecting model inputs, and tuning model parameters. However, the involvement of domain expertise will typically reduce the value of big data, as only a small subset of variables is used in model development. In addition, it is unlikely to discover novel knowledge, as model inputs and outputs are pre-defined. Training data refers to a set of observations where the input and output variables are both available. The quality of training data has a huge impact on the model reliability and robustness. It is worth mentioning that collecting high-quality training data can be costly, time-consuming and sometimes even not possible in practice.

By contrast, unsupervised analytics focus on discovering the intrinsic structure, correlations and associations in data. The success of implementing unsupervised analytics is not subject to the availability of training data, as there is no discrimination between inputs and outputs. The prominent advantage of unsupervised analytics is the ability to discover previously unknown knowledge [4, 57]. Supervised analytics adopts a backward approach in data analysis, which means the mining target (e.g., the model output) is pre-defined. Unsupervised analytics adopts a forward approach in data analysis. All the data are taken as inputs and the mining target is not explicitly defined. The ultimate goal is to reveal interesting relationships in data, if any. In such

a case, the value of big data can be best realized and the knowledge discovered might be valuable for practical applications.

This research emphasizes on the research and applications of unsupervised analytics in mining big building operational data. Considering the typical formats and types of building operational data, some of the most promising techniques are introduced as follows.

## **2.2 Clustering analysis**

Clustering analysis aims to group a set of observations into several clusters by maximizing the within-cluster while minimizing the between-cluster similarities. It is useful in revealing the underlying data structure and therefore, is usually applied for exploratory data analysis. Clustering analysis is typically used to analyze data stored in a single two-dimensional data table, where each observation is represented as a row and variables are represented as columns.

Clustering analysis involves five main tasks, i.e., feature generation, selection of proximity measures, clustering, data abstraction, and performance assessment [5].

Feature generation refers to the process of selecting or extracting features as inputs for clustering analysis. Proximity measures are then selected to evaluate the similarities among observations. The most widely used proximity measure for quantitative data is the Minkowski distance, i.e., denoted as  $(\sum_{i=1}^p |X_i - Y_i|^q)^{1/q}$ , where  $p$  is the number of variables and  $q$  is the order for distance calculation, e.g., the Euclidean distance is used when  $q$  is 2. Other popular proximity measures include the Mahalanobis distance,

Pearson correlation and Cosine similarity. The third task is to use a specific algorithm to cluster the data. Once the clustering is performed, data abstraction is performed to yield a compact description of each cluster. This is usually done for result interpretation, as there is no label to describe the data characteristics in each cluster. The quality of clustering can be evaluated using either the external (e.g., purity and F-measure) or internal (e.g., Davies-Bouldin index and Dunn index) methods [6].

Clustering algorithms are generally classified into four categories, i.e., partition-based, hierarchical-based, density-based and grid-based algorithms [7]. Partition-based algorithms adopt an iterative approach to arranging observations into several clusters. Each observation is assigned according to its distance to different cluster centers. The most classical partitioning-based algorithms are the  $k$ -means and  $k$ -medoids algorithms. The most important parameter is the cluster number  $k$ , which can be quite difficult to determine when prior knowledge is limited. The main advantages are the ease of implementation and high efficiency in clustering large data sets. However, the intrinsic assumptions indicate that reliable clustering results can only be obtained when the underlying clusters are spherical, evenly sized and uniformly distributed in data space [8]. The performance can be negatively affected due to the existence of outliers. Besides, the clustering results obtained are stochastic due to the random initialization of clustering centers or centroids.

By contrast, hierarchical clustering is deterministic, fairly robust to outliers and capable of revealing non-spherical clusters. The hierarchical clustering algorithms can be further divided into two groups, i.e., agglomerative and divisive. Agglomerative

algorithms adopt a bottom-up approach in constructing clusters, i.e., each observation is regarded as a cluster at the beginning and then recursively merged based on pairwise similarities. Divisive algorithms start with a single cluster containing all the observations and keep splitting until singleton clusters are derived. The main drawback of hierarchical-based algorithms is their high computation load due to the calculation of pairwise proximities [9].

Rather than considering the distance between observations, density-based algorithms group observations into regions with higher density and separated by regions with lower density. One representative density-based clustering algorithm is the DBSCAN (i.e., density-based spatial clustering of applications with noise). It requires users to specify a radius for defining neighborhood and a minimum number of observations for identifying clusters. Parameter setting can be time-consuming and is usually done by trial and error. Density-based algorithms have the ability to identify outliers and clusters with arbitrary shapes (e.g., ellipsoidal and spiral), sizes and densities [10].

Another type of clustering algorithms is grid-based. The basic idea is to define multi-resolution grids in data space and then perform clustering based on grids instead of the data observations themselves. It is computationally efficient and can handle clusters with arbitrary shapes. Popular algorithms include STING, WaveCluster and CLIQUE [10].

Recent development in clustering analysis focuses on analyzing large-scale and high-dimensional data. Two promising topics are ensemble clustering and subspace clustering. Ensemble clustering borrows the idea of ensemble predictive models and



utilizes a number of clustering outcomes to ensure the result robustness. It has proved to be good alternatives when handling complicated and high-dimensional data sets [11]. Subspace clustering algorithms adopts a different approach to tackling high dimensionality. These algorithms firstly identify influential variables, based on which clustering is performed. It is claimed that subspace clustering can alleviate the noisy effect in high-dimensional data [12].

### **2.3 Association rule mining**

Association rule mining (ARM) is one of the most powerful unsupervised analytics. It has the ability to extract associations among variables and express knowledge discovered in a rule format. An association rule is defined as  $A \rightarrow B$ , where  $A$  and  $B$  are two item sets and  $A \cap B = \emptyset$ .  $A$  and  $B$  are called antecedent and consequence respectively. It states that if  $A$  happens, then  $B$  will also happen. Besides its original application in the retailing industry for analyzing customer purchasing behaviors, ARM has been successfully applied in various industries, such as healthcare and financial industries [4].

There are two key steps in conventional ARM algorithms (e.g., Apriori), i.e., discovering frequent item sets and generating association rules [13]. Two parameters, i.e., support and confidence, are pre-defined for deriving association rules. Support is the joint probability of  $A$  and  $B$  both happening and denoted as  $P(A \text{ and } B)$ . Confidence is the conditional probability of  $B$  given  $A$  and denoted as  $P(B/A)$ . An association rule is derived only if it has a support and confidence no less than the

minimum thresholds defined. A larger support threshold will lead to the discovery of more frequent associations. The confidence threshold is usually maintained above 80% to ensure the usefulness of association rules. Another useful statistic is lift, which is defined as  $\frac{Confidence(A \rightarrow B)}{Support(B)}$ . It is commonly used for selecting potentially useful rules.

A lift value larger or smaller than 1 indicates that the occurrence of antecedent will positively or negatively affect the occurrence of consequence. A lift value of 1 indicates that the antecedent and consequence are completely independent and therefore, the rule is generated by coincidence.

Conventional ARM algorithms, such as Apriori and FP-growth, are used to discover frequent patterns from categorical and cross-sectional data [14]. Studies have been carried out to extend the power of ARM in analyzing more complex data and extracting more complicated types of associations. Three branches of ARM appear to be particularly promising in mining building operational data, i.e., quantitative association rule mining (QARM), temporal association rule mining (TARM) and gradual association rule mining (GARM).

QARM can extract associations from mixed type of data (i.e., containing both categorical and numeric data). A naïve solution is to transform numeric data into categorical data based on domain expertise or simple discretization methods (e.g., equal frequency or equal width binning) [15]. Nevertheless, the information loss can be too high to guarantee the quality of associations discovered. Advanced algorithms have been developed to better perform the data discretization using clustering [16],

evolutionary [17] and fuzzy [18] approaches. These algorithms are especially useful when prior knowledge on the underlying data characteristics is limited. TARM has the ability to extract associations under certain temporal constraints. The format of temporal association rule is denoted as  $A \xrightarrow{T} B$ , which means that if  $A$  happens,  $B$  will happen within  $T$  time steps. Various algorithms, such as RuleGrowth [19] and ERMiner [20], have been developed to mine temporal association rules. The knowledge discovered by TARM can be of great value when temporal interactions are of concerns. GARM aims to reveal a special kind of associations in numeric data, i.e., gradual dependency [21]. Gradual association rules are represented as “the more/less  $A \rightarrow$  the more/less  $B$ ”, which in essence describe the co-variations in numeric data. Advanced algorithms have been developed to ensure the mining efficiency of gradual dependency in large data, such as the GRITE [22] and GEP [23].

## 2.4 Motif discovery

Motif discovery, or sequence motif discovery, aims to extract frequently occurring subsequences in time series. The knowledge discovered is represented as collections of subsequences with similar patterns. It serves as foundations for many time series data mining tasks, such as temporal association rule mining, discord (i.e., atypical subsequences) detection and time series classification [24].

The basic version of motif discovery algorithms aims to discover frequently occurring subsequences in univariate time series. A univariate time series is firstly segmented into different subsequences based on a user-defined window size. The similarities

between different pairs of subsequences are calculated using certain proximity measures. Subsequences with high similarities are then identified as motifs. Considering that time series data usually of great length, the computation load is usually very heavy using the exhaustive search methods. Researchers have developed solutions to improve the computational performance from two perspectives, i.e., dimensionality reduction and efficient pattern matching. One of the most widely used methods is based on the symbolic aggregate approximation (SAX) and the random projection algorithm [25, 26]. The algorithm firstly transforms the univariate time series into a sequence of symbols with the aim of dimensionality reduction. The random projection method is then integrated to achieve efficient pattern matching. The algorithm can greatly reduce the computation load by approximating exact matching results through a few iterations. Advanced algorithms have been developed to discover motifs in multivariate time series. As examples, the principal component analysis and density estimation-based algorithms have been developed to identify synchronous multivariate motifs [27, 28]. These algorithms are subject to a strong assumption that multivariate motifs only consist of univariate motifs taking place during the same time period. To enable the discovery of both synchronous and non-synchronous multivariate motifs, some complicated algorithms have been proposed [29].

## **2.5 Unsupervised anomaly detection**

Anomaly detection intends to identify observations that deviate from expected or

normal behaviors [30]. Supervised anomaly detection can be achieved using classification techniques, e.g., developing a classification model to decide whether an observation is “normal” or “abnormal”. Considering that labels are seldom available in practice, unsupervised anomaly detection is more flexible to use. In general, data anomalies can be categorized into two kinds, i.e., (1) global anomalies, which are observations that are very different from the majority of the data; (2) local anomalies, which are observations that are perceived as anomalies only when considering a certain part of the data. Fig. 1 serves as an illustration of these two kinds of anomalies. The red round points are global anomalies as they greatly deviate from the majority of the data. The green triangle point represents a local anomaly. It might be regarded as a normal observation considering the whole data space, but stands out as an anomaly when focusing on observations in the lower left corner.

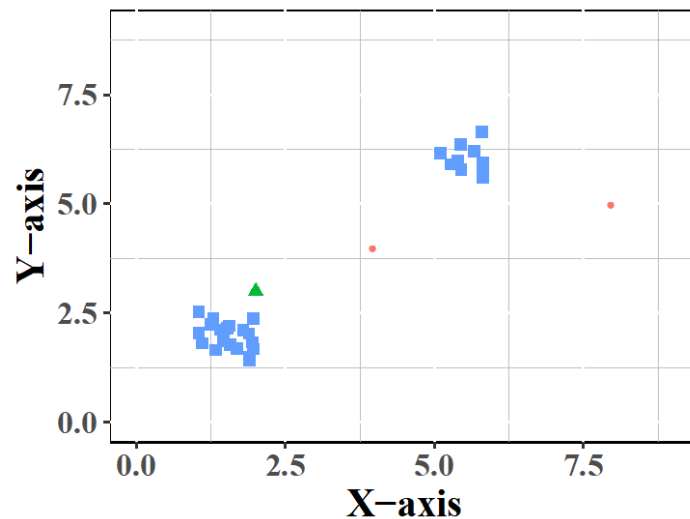


Fig.1 An illustration of global and local anomalies

The unsupervised anomaly detection methods can be classified into three groups, i.e., nearest neighbor-based methods, clustering-based methods, and statistical methods

[31]. The most classic nearest neighbor-based anomaly detection method is  $k$ -nearest neighbor. It can detect global anomalies by assigning an outlier score to each observation according to its distances to the top  $k$  nearest neighbors. Extensions of  $k$ -nearest neighbor have been developed to detect local anomalies, such as the local outlier factor (LOF), connectivity-based outlier factor (COF) and local correlation integral (LOCI) [31]. Clustering-based methods integrate clustering analysis and density estimation methods for anomaly detection [31]. For instance, the cluster-based local outlier factor (CBLOF) algorithm firstly applies  $k$ -means clustering to determine dense regions in data. The resulting cluster size is used as the density indicator. An anomaly score is assigned for each observation considering its distance to the cluster center and the cluster size. Statistical anomaly detection methods typically assume a data distribution and then use statistical criteria to identify anomalies. One prominent example is the generalized extreme studentized deviate (GESD) method. It can be used to detect one or more anomalies in univariate data. The method relies on the concept of  $t$ -distribution and hypothesis testing. A parameter  $u$ , which specifies the suspected number of anomalies, has to be pre-defined. Each observation is assigned with an anomaly score considering its distance to the mean and the intrinsic variations in data. Statistical methods have obvious limitations, as real-world data may not fit a certain statistical distribution and the problem become worse as data dimensionality increases [32, 33].

The recent development in unsupervised anomaly detection focuses on detecting anomalies from high-dimensional data. There are three general approaches, i.e., (1)

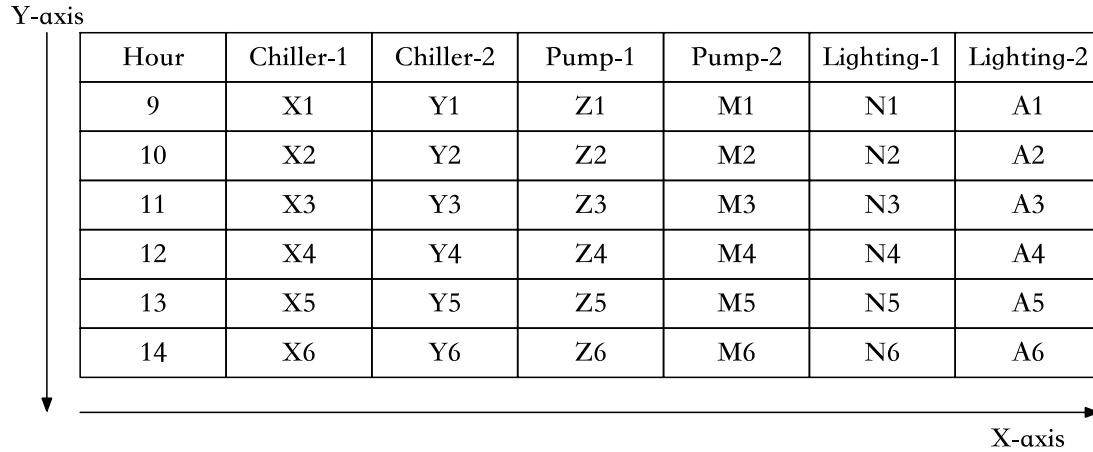
using more robust similarity metrics; (2) subspace anomaly detection; (3) ensemble anomaly detection. The first approach argues that distance metrics will become less meaningful in high-dimensional data (i.e., due to the curse of dimensionality) and therefore, using more robust similarity metrics can enhance the anomaly detection performance. An example of this approach is the angle-based outlier detection (ABOD) algorithm [34]. ABOD assigns a score to each observation based on the variance of all angles formed by this point and all the other pairs of data. The rationale behind is that if this observation is an anomaly, all the other observations should be lying remotely in a similar direction and hence, the variance of angles is small. If this observation is not an anomaly, it should be surrounded closely by the other data and hence, the angle variance is large. The second approach performs anomaly detection in subspace and thereby, reducing the masking effects of irrelevant variables. The general step is to firstly identify relevant variables for anomaly detection and then apply distance measures for anomaly detection [35]. The third approach integrates the concept of ensemble learning to enhance anomaly detection performance in high-dimensional space. For instance, users could calculate anomaly score  $v$  times, each using a randomly selected subset of variables or a certain anomaly detection algorithm [36]. For each observation, the final score is a certain combination of  $v$  scores (e.g., mean, maximum or minimum). Such ensemble methods could enhance the result reliability and robustness [37].

### **3. Data and knowledge at building operation stage**

### 3.1 Types and formats of building operational data

Building operational data are typically recorded in a two-dimensional structured data table, where each column represents a variable and each row stores the measurements or signals at the same time step. Fig. 2 presents an example data table, which records the power consumptions of chillers, pumps and lighting system at different time steps. There are four types of commonly recorded data during building operations, i.e., time data, energy consumption data, system operating parameters, and environmental data. Time data refer to the time and date when measurements are taken, such as *Year*, *Month*, *Day*, *Hour*, *Minute*, *Second* and *Day Type* (i.e., Monday to Sunday). Energy consumption data refers to power consumptions of various services systems and their components, cooling and heating loads, natural gas consumptions and etc. System operating parameters provide detailed descriptions about operating conditions. The main focus of BAS is the heating, ventilation and air-conditioning (HVAC) system, as it accounts for the largest energy use in commercial buildings and has great energy saving potentials [38]. Typically recorded operating parameters include the temperature, flow rate and pressure of chilled and condenser water distribution loops, the frequency of motor drives, on-off status of different components, and all kinds of set-points for HVAC controls and management. Environmental data refers to the indoor or outdoor built environment, such as the dry-bulb temperature, relative humidity, wind speed and solar radiation.





Hour	Chiller-1	Chiller-2	Pump-1	Pump-2	Lighting-1	Lighting-2
9	X1	Y1	Z1	M1	N1	A1
10	X2	Y2	Z2	M2	N2	A2
11	X3	Y3	Z3	M3	N3	A3
12	X4	Y4	Z4	M4	N4	A4
13	X5	Y5	Z5	M5	N5	A5
14	X6	Y6	Z6	M6	N6	A6

Fig.2 Typical formats for building operational data

The majority of the data are recorded as numeric data. However, some of them should be treated as categorical despite of their numeric appearance. For instance, the time variable *Day* ranges from 1 to 31. It should be treated as categorical, as differences in numeric values do not indicate one particular *Day* is more or less significant than another. Distinctions should be made to avoid misleading analysis results.

### 3.2 Types of knowledge hidden in building operational data

According to the typical formats of building operational data, there are two main types of knowledge to be discovered, i.e., (1) cross-sectional or static knowledge, and (2) temporal or dynamic knowledge.

Cross-sectional knowledge can be discovered when analyzing building operational data as cross-sectional data (i.e., along the *X*-axis as shown in Fig. 2), where each row is treated as an independent observation. In such a case, the knowledge discovered is static as the temporal dependencies between different rows are neglected. Cross-sectional knowledge is the basic knowledge type hidden in building operational

data. The majority of data analytics are designed to extract cross-sectional knowledge, such as the Apriori algorithm and the  $k$ -means clustering algorithm. Cross-sectional knowledge is useful for the identification of frequent interactions among building components, typical and atypical operating conditions and etc. By contrast, capturing temporal dependencies is more challenging. It can be achieved by mining building operational along both the  $X$ - and  $Y$ -axes of the two-dimensional data table. Temporal knowledge is very useful for characterizing dynamics in building operations. The insights obtained can be used for developing dynamic solutions for building optimal control, fault detection and diagnosis.

#### **4. Applications of unsupervised data analytics for building energy efficiency enhancement**

As illustrated in Fig. 3, the general process of analyzing building operational data includes four main phases, i.e., data preprocessing, exploratory data analysis, knowledge discovery and post-mining. Data preprocessing mainly includes data cleaning and data transformation. Data cleaning refers to the process of integrating data from different sources, handling missing values and removing outliers. Data transformation aims to prepare data into suitable format for DM algorithms, e.g., using principal component analysis to reduce dimensionality and applying discretization methods to transform continuous numeric variables into categorical variables. The second phase is exploratory data analysis, which is a rather a broad topic. It aims to gain insights into data characteristics in terms of data structures,

variable types and distributions. The ultimate goal of exploratory data analysis is to provide clues for further in-depth analysis. The third phase, i.e., knowledge mining, performs knowledge discovery using either supervised or unsupervised analytics. The knowledge obtained at its raw form can be redundant and difficult to interpret. Hence, the final phase, i.e., post-mining, is designed for the ease of practical applications. Post-mining methods are developed for knowledge interpretation (i.e., express raw knowledge using interpretable representations) and knowledge selection (i.e., screen out valid and potentially useful knowledge).

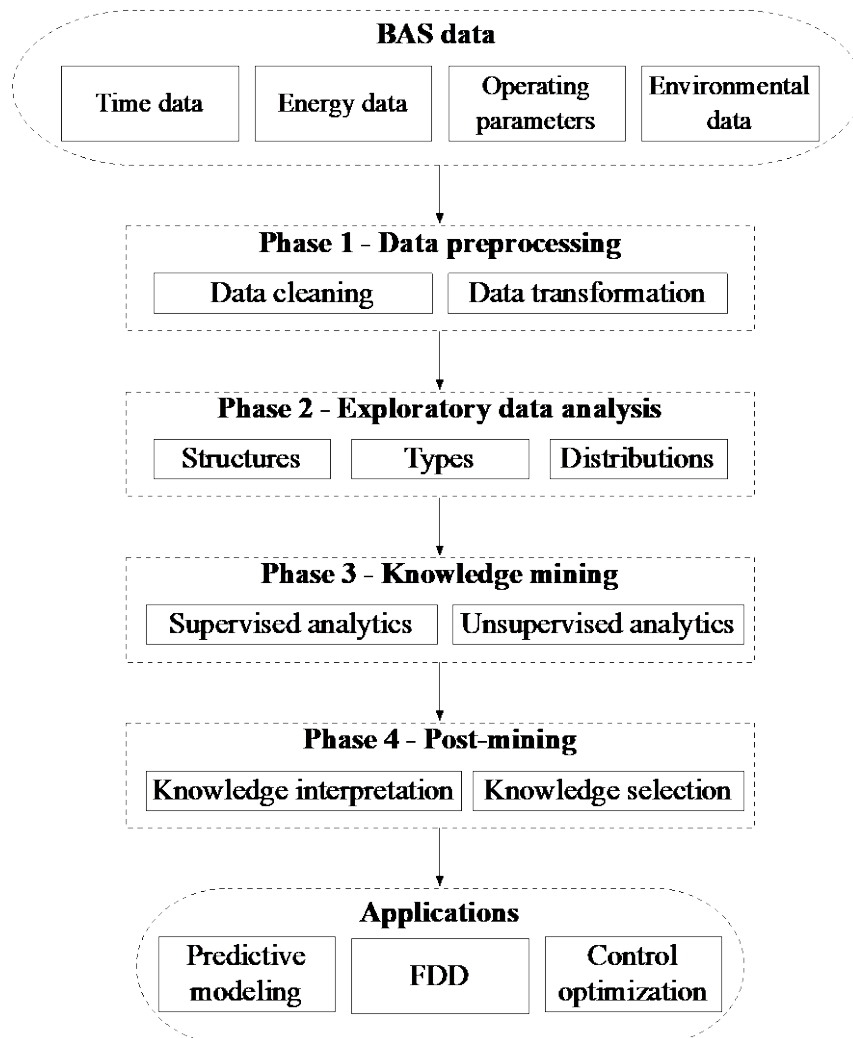


Fig. 3 The general process for analyzing massive building operational data

Previous studies have demonstrated the value of unsupervised data analytics in analyzing building operational data. Unsupervised analytics are mainly used in two phases, i.e., exploratory data analysis and knowledge mining. This section focuses on the application of unsupervised analytics in the following aspects: (1) identifying static building operation patterns; (2) identifying temporal building operation patterns; (3) detecting anomalies in building operations; (4) identifying occupant behavioral patterns.

#### 4.1 Identifying static operation patterns

As introduced in Section 3, cross-sectional or static knowledge is the basic type of knowledge that can be extracted from building operational data. The static knowledge discovered from building operational data has been applied for identifying the intrinsic data groupings, frequently occurring operating conditions, and significant interactive relationships among building variables [39, 40]. To summarize, the most commonly used unsupervised analytics are clustering analysis and association rule mining. The static knowledge discovered is represented as data clusters and association rules respectively. Table-1 serves as an overview of the publications reviewed in this section.

Table-1 Publications related to the identification of static operation patterns

Year	Applications	Unsupervised analytics	Algorithms	Tools	Refs
2010	Heating system	Clustering analysis	<i>k</i> -means	<i>MATLAB</i>	[41]

	performance evaluation				
2011	HVAC system performance evaluation	Clustering analysis	Agglomerative hierarchical clustering	Not specified	[42]
2012	Chiller system performance assessment	Clustering analysis	Two-step clustering	<i>SPSS</i>	[43]
2015	Building system performance evaluation	Clustering analysis & QARM	<i>k</i> -means, <i>k</i> -medoids, fuzzy c-means, EWKM	<i>R</i>	[44]
2016	Building energy waste identification	Clustering analysis	DBSCAN	Not specified	[45]
2012	Chiller system performance evaluation	Clustering analysis	Two-step clustering	<i>SPSS</i>	[46]
2017	Variable refrigerant flow system performance evaluation	Clustering analysis & ARM	<i>k</i> -means, Apriori	<i>R</i>	[47]
2017	District heating system	Clustering analysis &	<i>k</i> -means, <i>k</i> -medoids,	<i>R</i>	[48]

	performance evaluation	ARM	hierarchical clustering, Apriori		
2017	Building energy usage pattern identification	Clustering analysis	<i>k</i> -shaped	<i>R</i>	[49]
2014	Building energy usage prediction	Clustering analysis	Entropy-weighted <i>k</i> -means	<i>R</i>	[50]
2014	HVAC energy usage prediction	Clustering analysis	<i>k</i> -means	Not specified	[51]
2011	Building energy load management	Clustering analysis	Hierarchical clustering	Not specified	[52]
2015	Building energy load prediction	Clustering analysis	<i>k</i> -means	Not specified	[53]
2012	HVAC system performance evaluation	ARM	FP-growth	<i>RapidMiner</i>	[54]
2013	Lighting energy waste identification	ARM	Apriori	Not specified	[55]

#### 4.1.1 Static knowledge discovery based on clustering analysis

Conventional clustering analysis algorithms, such as *k*-means and *k*-medoids, are capable of grouping observations into different clusters according to their similarities.

The observations assigned in the same cluster can be regarded as measurements with

similar operating conditions. The resulting cluster center or centroid can be used as representations of typical building operating patterns.

Previous studies have applied clustering analysis for identifying typical operation patterns at three levels, i.e., building-level [41], system-level [42] and component level [43]. The key of applying clustering analysis for typical operation pattern identification is clustering input selection. For instance, building-level operation characteristics can be well described by building energy consumption, indoor and outdoor environment and therefore, these variables are usually selected as clustering inputs [44]. Howard et al. developed an automated method to identify energy efficiency opportunities using the whole building energy consumption data [45]. The DBSCAN algorithm was utilized to cluster the residuals generated by robust regression model, based on which groups of days exhibiting various energy consumption patterns were discovered. System-level and component-level operation patterns can be identified based on operation inputs and outputs. As an example, the input of HVAC system is the electricity power and operating parameters, while the output is the cooling load. Typical operation patterns of HVAC system can therefore be identified using these variables as clustering inputs, e.g., the power consumptions of chillers, pumps, cooling towers and the building cooling load [46, 47].

Once typical operation patterns are identified, further analysis can be performed on each cluster separately to enhance the reliability and sensitivity of knowledge discovered. Therefore, clustering analysis has been applied as a tool for data preparation or exploratory data analysis. The knowledge discovered by clustering

analysis serves as foundations for further in-depth analysis, such as operation pattern identification and building energy prediction [48, 49]. Fan et al. adopted the entropy-weighted *k*-means (EWKM) clustering algorithm to identify typical daily electricity consumption patterns of commercial buildings [50]. Twelve statistical features were extracted to represent daily energy consumption profiles. EWKM successfully identified two clusters, each containing observations from weekdays and weekends respectively. Predictive models on daily energy consumptions were developed separately for each cluster, resulting in much more accurate predictions. Similarly, Tang et al. applied *k*-means algorithm to cluster the whole data before developing predictive models [51]. It was claimed that such method was able to reduce the prediction error and computational load. Jota et al. developed a 4-step methodology to forecast the building electricity consumption [52]. Hierarchical clustering was applied to identify typical operation patterns. The results showed that the proposed method could predict the energy consumption and maximal demand with high efficiency. Hsu compared the building energy prediction performance using cluster-wise regression and two-stage process of *k*-means and model-based clustering [53]. The results showed that there is a tradeoff between prediction accuracy and cluster stability.

#### **4.1.2 Static knowledge discovery based on association rule mining**

Conventional association rule mining (ARM) algorithms, such as Apriori, has been applied for identifying static relationships among building operational data. The



knowledge is represented as “IF-THEN” rules. Yu et al. applied ARM to discover significant associations in HVAC system operations [54]. The association rules obtained were successfully used to detect equipment faults and identify energy waste conditions. Cabrera and Zareipour used ARM to identify the lighting energy waste patterns in educational institutes [55]. The data set contained 7 variables, i.e., season, time, day type (i.e., weekdays or weekends), occupancy status, event, day of week (i.e., Monday to Sunday) and waste status. Association rules were derived to describe the relationships between energy waste and the other variables. The knowledge discovered was used to regulate the lighting energy use. It was shown that up to 70% of the lighting energy could be saved. Li et al. adopted ARM to extract rules describing the operation patterns of variable refrigerant flow systems [47]. Energy consumption patterns related to different part-load ratios, refrigerant charge levels and cooling conditions were successfully discovered. Xue et al. adopted clustering analysis and ARM to identify operation patterns in district heating stations [48]. Three clustering algorithms, i.e., *k*-means, *k*-medoids and hierarchical clustering, were applied to identify seasonal and daily operation patterns. A number of meaningful association rules were discovered using the Apriori algorithm, providing clues on energy-inefficient operations.

One major limitation of conventional ARM algorithms is that they can only handle categorical variables. Building operational data are mostly numeric and therefore, data discretization becomes a necessary step in mining building operational data. While some straightforward discretization methods, such as the equal-frequency and

equal-width binning methods, are easy for implementation, they may result in improper discretization results with high information loss. Admittedly, manually discretizing each numeric variable based on domain expertise can reduce the information loss, but is time-consuming and not easy for generalization. The quantitative association rule mining (QARM) is developed to tackle this dilemma. Advanced QARM algorithms can automatically identify the most suitable intervals for data discretization while considering the quality of the association rules discovered. It has been employed in mining quantitative associations in building operational data [44]. The quantitative associations discovered can bring more insights into building operations, as the discretization interval is identified through a data-driven approach.

The main challenge in applying ARM is that the number of association rules discovered is usually large, especially when the support and confidence thresholds are set low. Post-mining methods therefore should be developed for the efficient and effective utilization of knowledge discovered. The most commonly used solution is to construct rule engines based on association rules discovered. The general idea is to use the rule engine to scan the building operational data. The abnormality of each observation can be evaluated based on the number and quality of rules it violates (i.e., a rule is violated if an observation meets the antecedent but fails to meet the consequence). An example was given in [44], where an abnormality degree was calculated for each observation based on rules being violated. The method was proved to be useful in identifying atypical operations.

## 4.2 Identifying temporal operation patterns

Temporal knowledge can be discovered when treating building operational data as multivariate time series. The intrinsic challenges in discovering temporal knowledge from building operational data are two-fold. Firstly, time series are usually of great length, which dramatically increase the complexity in computation. Secondly, the algorithms are usually more complicated and less known to building professionals. As a result, the research related to temporal knowledge discovery in building operational data is rather limited. To summarize, three types of unsupervised analytics have been found in previous studies, i.e., (1) symbolic approximate aggregation (SAX); (2) clustering analysis; (3) motif discovery; (4) temporal association rule mining. Table-2 provides a summary on the publications reviewed in this section.

The SAX method is adopted to tackle the challenge of high dimensionality in time series data, i.e., a time series of length  $n$  will be transformed into a sequence of  $m$  symbol, where  $m$  is much smaller than  $n$ . The SAX method gains its popularity as the information loss in data transformation is bounded [25]. In addition, the SAX representations are compatible with many distance-based mining algorithms. It can be used to identify frequent and infrequent sequential patterns in building operations. Miller et al. developed an automated filter to find infrequent daily patterns in building operations [57]. The SAX method was applied for data transformation. The most infrequent sequential patterns, or discords, were filtered out for detailed inspection. The method was applied to two case studies and the results confirmed its capability of

finding discords in time series data.

Clustering analysis has also been applied to identify temporal patterns in time series data. A conventional approach is to prepare the time series into subsequences, based on which clustering analysis is performed to group subsequences into different clusters [56]. The knowledge discovered is especially useful for load profiling and customer classification [58]. One essential application of load profiling is to identify typical patterns of individual building customers. Accurate grouping was achieved by clustering the building energy usage data using *k*-means algorithm [59]. The performance of various clustering algorithms has been compared and validated for electricity load profiling [60]. Based on the results of load profiling, customer classification can be performed. Figueiredo et al. used *k*-means and labeled account information to classify customer behaviors [61]. As a general solution, Vale et al. proposed a DM-based framework for characterizing building customers [62].

The identification of frequent sequential patterns is also referred as motif discovery. It was firstly applied in [56] to extract sequential patterns in chiller power consumptions. A number of sustainable metrics (e.g., part-load ratios and average power consumption) were then calculated for each motif, based on which sustainability characterization of chiller operations was achieved. Motif discovery is mainly applied for mining univariate power consumption time series in building operational data [63, 64]. The most widely used algorithm is based on the concept of SAX representations and random projection [25]. Some advanced algorithms have been developed to discover multivariate motifs. Fan et al. have applied a multivariate algorithm to

identify multivariate motifs among power consumptions of different services systems, such as chillers, cooling towers, and lighting systems [65]. The knowledge discovered is useful to describe dynamics in building operations and temporal interactions among different variables. Kalluri et al. developed a time series data mining approach to extracting temporal operating patterns in office plug load appliances [66]. The SAX method was firstly applied to transform plug load into symbols. The motifs were then discovered through grammar induction. The knowledge discovered were proved to be useful for characterizing and classifying appliance signatures. In addition, several studies have been performed to identify the embedded sub-motifs in time series data, based on which building power disaggregation can be achieved [67-69]. Aiad and Lee developed an unsupervised method to breakdown the total household power consumptions into power consumed by individual appliances [70]. The Factorial Hidden Markov Models (FHMMs) were used to represent the interactions among different devices. The performance was validated through a public data set and accurate disaggregation results could be obtained. Ulmeanu et al. adopted HMMs to analyze hourly-recorded time series of building thermal loads [71]. The research results showed that various temporal patterns, each corresponding to different seasonal effects and occupancy behaviors, could be successful identified.

Another type of temporal knowledge can be expressed as temporal association rules. Temporal association rules have a format of  $A \xrightarrow{T} B$ , which states that if  $A$  happens,  $B$  will happen  $T$  time steps later. The knowledge discovered by temporal association rule mining (TARM) can be very helpful for understanding the dynamics in building

operations and identify anomalies in temporal space. However, studies on this topic are rather limited. Fan et al. applied the TRuleGrowth algorithm to extract temporal associations in chiller operations [65]. The status of chiller operation was represented as two time series, i.e., one represents the power consumption level and the other one represents the trending information. The temporal associations discovered were found to be useful for identifying time lags in status transitions and anomalies in temporal space.

Table-2 Publications related to the identification of temporal operation patterns

Year	Applications	Unsupervised analytics	Algorithms	Tools	Refs
2009	Chiller operation pattern identification	Clustering analysis & Motif discovery	$k$ -means, Apriori-based motif mining	Not specified	[56]
2015	Building energy usage pattern identification	Clustering analysis & Motif discovery	SAX, $k$ -means	<i>VizTree</i>	[57]
2015	Building energy usage pattern identification	Clustering analysis	$k$ -means	Not specified	[59]
2013	Building energy usage pattern identification	Clustering analysis	Fuzzy clustering	Not specified	[60]
2005	Building energy usage pattern identification	Clustering analysis	$k$ -means, Self-organizing map	Not specified	[61]

2009	Building energy usage pattern identification	Clustering analysis	Unspecified	Not specified	[62]
2011	Chiller operation pattern identification	Clustering analysis & Motif discovery	<i>k</i> -means, Apriori-based motif mining, agglomerative hierarchical clustering	Not specified	[63]
2013	Building energy usage disaggregation	Clustering analysis & Motif discovery	Hierarchical clustering, Apriori-based motif mining	Not specified	[64]
2015	Building energy usage pattern identification	Motif discovery & TARM	SAX, TRuleGrowth, Random projection	<i>R</i> ; <i>SPMF</i>	[65]
2016	Plug load pattern identification	Motif discovery	SAX	<i>GrammarViz</i>  <i>2.0</i>	[66]
2016	Building power usage disaggregation	Motif discovery	Factorial Hidden Markov Models	Not specified	[70]
2017	Building power usage	Motif discovery	Hidden Markov	Not Specified	[71]

	disaggregation		Models		
--	----------------	--	--------	--	--

### 4.3 Detecting anomalies in building operations

Anomalies in building operations include faults in building operations and atypical or infrequent operating conditions. Actually, the identification of frequent operation patterns can be used as foundations for anomaly detection. For instance, given a frequent association expressed as  $A \rightarrow B$ , anomalies can be detected if an observation meets the antecedent but fails to meet the consequence. Nevertheless, post-mining methods are needed to realize this functionality, which can be tedious and time-consuming. This section focuses on the unsupervised data analytics that are solely designed for anomaly detection. The publication reviewed in this section is listed in Table-3.

Statistical methods and clustering analysis are found to be the main tools used for detecting anomalies in building operational data. Statistical methods are primarily applied for detecting anomalies in univariate data, e.g., the total building power consumption [72]. Seem developed a generalized extreme studentized deviate (GESD)-based method to evaluate the abnormality degree of building electricity consumption [73]. The method firstly calculated summarizing statistics (e.g., mean and maximum) from daily electricity consumptions as features. The GESD method was then applied to evaluate abnormality degree of each daily profile. The method was proved to be computationally efficient and valid through field tests. Similarly, Lin and Claridge developed a statistical method to detect abnormal building energy



consumption based on temperature [74]. The abnormal operations were detected when the deviation between measured and simulated consumption was greater than one standard deviation of the residuals in the reference data. The performance was validated through simulation tests.

Clustering analysis can also be used to facilitate the task of anomaly detection [75]. It has been used in two ways. The first is to identify data clusters, based on which statistical methods are then used to detect anomalies in each data cluster. The second is to take advantage of the unique power of advanced clustering algorithms (e.g., DBSCAN) for the direct identification of anomalies [76]. Capozzoli et al. developed two approaches for detecting anomalies in building energy consumptions using statistical and clustering methods [77]. The first was to identify the intrinsic data structure using *k*-means clustering. The GESD method was then applied to detect anomalies in each cluster. The second was based on the DBSCAN algorithm alone, as it has the ability group outliers as a separate cluster. The research results showed that the DBSCAN algorithm could be very promising for identifying anomalies in multivariate data, given that the parameters were set properly. The study carried out by Jalori and Reddy further validated the usefulness of the DBSCAN algorithm in identifying anomalies in building energy consumption data [78]. Rather than using aggregated daily building energy consumption, this study adopts a 24-dimension vector (representing hourly energy usage in a day) as inputs for clustering analysis. The research results showed that the method could successfully identify daily anomalies together with typical operating schedules. As an extension to this work, the

knowledge discovered by clustering analysis was utilized to achieve accurate short-term building load forecasting [79].

The recent development in artificial intelligence and machine learning has provided new tools for unsupervised anomaly detection. As an example, Araya et al. proposed an ensemble framework to identify anomalies in building energy consumption data [80]. The autoencoder algorithm, which is an unsupervised version of artificial neural network, has been adopted as the key component in anomaly detection ensembles together with some predictive algorithms. The research results indicated that the performance was satisfactory in terms of true positive rate and false positive rate. Pena et al. proposed an intelligent rule-based approach to identifying anomalies in smart buildings [81]. A set of energy efficiency indicators were developed, based on which rules describing anomalies were successfully derived.

Table-3 Publications related to the identification of anomalies in building operations

Year	Applications	Unsupervised analytics	Algorithms	Tools	Refs
2010	Identifying daily anomalies in building energy data	Statistics	GESD	Not specified	[72]
2007	Identifying daily anomalies in building energy data	Statistics	GESD	Not specified	[73]
2015	Identifying point	Statistics	Standard	Not specified	[74]

	anomalies in building energy data		deviation-based algorithm		
2010	Identifying point anomalies in building energy data	Statistics & Clustering analysis	Standard deviation-based algorithm, K-nearest neighbor, Discrete time warping	Not specified	[75]
2013	Identifying point anomalies in building energy data	Statistics & Clustering analysis	GESD, <i>k</i> -means, DBSCAN	Not specified	[76]
2015	Identifying point anomalies in building energy data	Statistics & Clustering analysis	GESD, <i>k</i> -means, DBSCAN	Not specified	[77]
2015	Identifying daily anomalies in building energy data	Clustering analysis	DBSCAN	Not specified	[78]
2015	Identifying daily anomalies in building energy data	Clustering analysis	DBSCAN	Not specified	[79]
2017	Identifying contextual	Unsupervised learning	Autoencoder	<i>H2O</i> ; <i>R</i>	[80]

	anomalies in building energy data				
2016	Identifying point anomalies in building energy data	Rule-based methods	Unspecified	<i>SPSS</i>	[81]

#### 4.4 Identifying occupant behavioral patterns

Occupant behavior has a major impact on building energy consumptions. It has been identified as the key to bridge the gap between predicted and actual building energy consumptions [82]. However, identifying the influence of occupant behavior on building energy efficiency can be very challenging. The reasons are two-fold. Firstly, occupant behavior has a stochastic, diverse and complex nature. The acquisition of an accurate description on occupant behavior is usually impractical. Secondly, building energy consumptions are subject to the influences of many factors, such as surrounding climate, building materials and economics. Hence, it can be very difficult to quantify the isolated effect of occupant behavior on building energy consumption.

Researchers have turned to unsupervised analytics for solutions. Yu et al. adopted *k*-mean clustering analysis to examine the effects of occupant behaviors [83]. The data mining software *WEKA* was used for analysis. A number of buildings were firstly divided into various groups based on four factors unrelated to occupant behaviors, i.e., climate, building-related characteristics, number of occupants, and building services systems. The energy differences among observations in the same cluster were treated

as the consequences of different occupant behaviors. Occupant behaviors were represented by the power consumptions of eight end-use loads, including HVAC, hot water supply, kitchen, lighting, refrigerator, amusement and information, housework and sanitary, and others. The method allowed researchers to identify the influence of occupant behaviors, based on which energy saving measures were developed. Capozzoli et al. developed a methodology to reduce the HVAC system energy consumption through occupancy pattern learning [84]. The *k*-means algorithm was used to identify occupancy patterns in different thermal zone. An optimized HVAC control strategy was then developed by grouping similar occupancy patterns into the same thermal zone. The simulation results showed that up to 14% of HVAC energy consumption could be reduced compared with the occupancy-independent operation schedule.

ARM has been found promising in revealing occupant behavioral patterns. Yu et al. applied ARM to analyze the power consumptions of different household appliances (such as air-conditioner, dishwasher, washing machines and microwave oven) and environmental variables (e.g., outdoor temperature and relative humidity) [85]. The analysis was carried out using software *WEKA*. The association rules obtained could be used to describe the interaction between occupant behavioral patterns and environment variables. The insights obtained can be used to provide detailed recommendations for reducing building energy consumption.

A number of studies have adopted Hidden Markov Models (HMMs) to develop unsupervised methods for occupancy detection and occupant behavior identification

[86-88]. As an example, Candanedo et al. proposed an unsupervised occupancy detection method based on HMMs [89]. The open-source program *R* and its package *depmixS4* were used as computation tools. The models were developed based on a set of features (e.g., CO<sub>2</sub>, temperature and humidity) and evaluated using a public data set. A case study was carried out to further validate the usefulness of the methodology on occupancy detection.

With the development in information technology, buildings are being equipped with advanced sensors to measure the occupancy (e.g., movement), human interaction with building envelope (e.g., windows and blinds) and services systems (e.g., HVAC and lighting). These data can be very helpful for an in-depth characterization of occupant behaviors [90]. As an example, D'Oca et al. proposed a DM-based framework to characterize the behavioral pattern on window usage [91]. Clustering analysis (using the *k*-means algorithm) and association rule mining (using the FP-growth algorithm) were employed as the main analytics for identifying the window-opening and window-closing patterns. The data mining software *RapidMiner* was used as mining tool. It was claimed that unsupervised DM techniques were proficient in highlighting behavioral patterns and could overcome the lack of personalization of statistical methods. Wang and Shao investigated the use of WiFi signals in identifying indoor occupancy patterns [92]. Association rule mining was applied to identify energy waste patterns. The method was validated through a case study and showed that up to 26% of lighting energy consumption can be reduced.

## **5. Challenges and future research directions**

### **5.1 Challenges**

The building industry is embracing the era of big data. Compared to supervised analytics, unsupervised analytics have the ability to better realize the value in big data. Through this literature review, it is found out that the application of unsupervised analytics in building operational data analysis is still an under-developed research area. Many challenges do exist to hinder the application of unsupervised data analytics in building energy management. From the authors' perspective, there are three main challenges, i.e., (1) poor data quality in building operational data; (2) knowledge gap between building engineers and unsupervised data analytics; (3) privacy issues.

A famous slogan in computer science is “garbage in, garbage out”, which means that if the input data is of poor quality, the results obtained will also be of little use. Currently, the building operational data quality is usually poor, containing a large amount of missing values, imprecise or dead readings. In addition, the data collection interval for some buildings can be too large to extract meaningful insights. For instance, the energy use of residential buildings is usually recorded monthly or annually from utility companies. In practice, it is rather difficult to develop generic solutions to improve the building operational data quality due to the variations in individual operating behaviors and operating budgets. One possible way to alleviate this problem is to create an information-sharing mechanism among different buildings for data quality cross-validation. However, this may cause some concerns over

information security and privacy issues.

The second challenge is the widespread knowledge gap between building engineers and unsupervised data analytics. Nowadays, building engineers mainly adopt physics and engineering experience to tackle problems in building operations. The lack of knowledge in data analytics makes them unwilling to turn to data-driven methods for solutions. Future building engineers should have the basic knowledge of unsupervised analytics in order to understand the data analysis report generated by the BAS system or even develop customized analysis tools for building energy management.

Thirdly, privacy issues have not been properly addressed in the building field, which negatively affects the practical applications and academic research of data-driven methods for building energy efficiency enhancement. Individual privacy is the main concern which forbids the widespread collection and usage of building occupant data. For instance, records on personal equipment power consumptions, indoor environment measurements on individual spaces, and video data collected by the building closed-circuit television system can be very useful for understanding building occupant behaviors and developing optimal building control strategies. However, this type of information is closely related to the individual privacy of building occupants. As a result, relevant data resources are still lacking and the scale of research related to building occupant behaviors is rather limited. In addition, it is noted that one of the best ways to improve the quality of academic studies and their practical values is to make the research reproducible and data available to public. From the authors' perspectives, reproducible research is the future trend, as it creates



a valuable information-sharing platform among researchers and building practitioners. Nevertheless, the privacy issues related to building owners and customers may hinder the development of reproducible research in the building field. So far, few studies and their results have been made reproducible due to the privacy concerns from building owners and customers. Industrial guidance is urgently needed to regulate the legal and ethical aspects of building data usage, so that in-depth and large-scale (e.g., at district- or city-scale) research can be performed.

## **5.2 Future research directions**

### **5.2.1 Generic post-mining methods**

Post-mining the knowledge discovered by unsupervised analytics can be time-consuming due to the following four issues: (1) knowledge amount; (2) knowledge redundancy; (3) knowledge characterization; (4) knowledge transformation. One essential advantage of unsupervised analytics is that they could achieve better data utilization. As a ‘dark-side’, the knowledge discovered usually has a large amount and a high redundancy. Taking association rule mining as an example, the number of rules obtained may increase rapidly with the decrease in support and confidence thresholds. In addition, it is very likely that redundant rules are generated through the mining process, i.e., Rule A is redundant if it is a super rule of Rule B but with the same or a smaller lift value. In such a case, the development of automatic or semi-automatic post-mining methods for knowledge selection is of great value.

Knowledge characterization refers to the process of generating labels or descriptions

for the knowledge discovered. For instance, clustering analysis can identify the intrinsic data groupings. However, it does not generate labels or descriptions for each cluster, making it difficult for users to understand the clustering results, especially when the number of clustering inputs is large. Further data exploration is needed for knowledge interpretation. So far, there is a lack of studies on improving the efficiency of knowledge characterization.

Another issue is knowledge transformation, which refers to the process of turning raw knowledge discovered into actionable measures for enhancing building energy efficiency. For instance, once the frequent or typical operation patterns are identified, further transformation (e.g., quantify the associations between different frequent patterns) is needed to make it applicable for practical applications (e.g., detecting abnormal operating patterns). Further studies should be performed to develop generic solutions in terms of different knowledge representations (e.g., association rules, cluster membership and anomalies)) and applications. This is a non-trivial task and requires users to have a deep understanding on both unsupervised analytics and building energy management.

### **5.2.2 Knowledge discovery in multi-relational databases**

It is noted that almost all the unsupervised analytics mentioned above are designed for mining data in a two-dimensional data table. With the advance in information technology, it can be foreseen that a diversity of information will be recorded and available for data analysis, such as the spatial and system-component affiliation

information. Consequently, more complex data structures, i.e., multi-relational database, will become prevalent for building operational data storage. Integrating multiple data tables into a single two-dimensional data table can be troublesome and sometimes not even possible. Therefore, advanced analytics and mining methodology should be developed to ensure the efficiency and effectiveness in analyzing multi-relational databases. Two general solutions are logic-based and graph-based approaches [93]. As summarized by Holder and Cook, logic-based approaches require a number of techniques for inducing a logical theory for data representation. By contrast, graph-based approaches can represent data and discover knowledge in a more flexible and visualized manner. Future research may focus on developing methods for analyzing building operational data stored in multi-relational databases. For instance, if the graph-based approach is selected, a thorough workflow should be designed, including generating graphs from conventional building operational data, selecting and modifying applicable graph-based mining algorithms, and developing methods for the post-mining of graph-based knowledge representations.

### **5.2.3 Knowledge discovery in unstructured data**

Building operations also generate massive amounts of unstructured data, such as texts generated from maintenance reports and video data generated from the closed-circuit television (CCTV) system. Previous studies mainly focused on analyzing structured data, while neglecting the value in unstructured building operational data. There are two general research directions in this area: (1) developing mining methodologies for

analyzing unstructured building operational data, including the general mining process and algorithms; (2) exploring their applications in improving building energy efficiency.

As an example, text data are being generated throughout the whole building life cycle, e.g., the specifications of building services systems or components, maintenance reports, questionnaires and social media of building occupants. Text mining aims to extract useful insights from text data. It has been utilized in various industries with success applications, such as analyzing sentiments from customer reviews and automatically generating labels for news [94]. The development of text mining-based methods enables new perspectives in enhancing building operational performance. For instance, one might perform sentiment analysis on the social media of building occupants and explores the underlying relationships with occupant behaviors and building energy performance.

Another promising topic is about video data mining [95]. The widespread use of the CCTV system has made video data an essential information source during building operations. The rapid development in data analytics has provided powerful tools to identify objects, characters and scenes in video data. The knowledge hidden in video data is very promising for monitoring indoor environment, detecting occupancy and identifying occupant behaviors [96-98]. Thanks to the development of advanced analytics such as deep learning, it is expected that more insights can be extracted from video data for building energy management.

## **6. Conclusions**

This paper provides a comprehensive review on the research of unsupervised data analytics in mining building operational data and its applications in improving building operational performance. Previous studies are reviewed based on the unsupervised analytics used, the resulting knowledge representations and their applications in building energy management. Compared to supervised analytics, unsupervised analytics do not require the availability of high-quality training data and are more promising in realizing the true value of big data.

Despite of some encouraging results, this field is still at its early stage. Among many challenges faced in the building field, the privacy issue related to building owners, building customers and building occupants is the most essential one. Guidance on the legal and ethical aspects of building data usage is urgently needed to provide resources for advanced and reproducible research. Several aspects deserving more research efforts are identified and proposed in this paper. One is to develop generic and effective post-mining methods for knowledge selection, interpretation and applications. Post-mining is the main challenge faced when applying unsupervised analytics for knowledge discovery. The development of semi-automated or fully automated post-mining methods can greatly reduce the complexity and therefore, deserves more in-depth research. Developing methodologies to analyze complex data is another promising research area in the building field. Previous studies focused on analyzing structured data that are stored in a single two-dimensional data table. The development in information and communication technologies has enabled the

collection of more complex data types and formats. In this regard, studies on how to extract meaningful knowledge from complex-structured (e.g., multi-relational data) and unstructured (e.g., text data and video data) building operational data should be performed. These research topics are very promising and can bring new perspectives in utilizing the information generated during building operations for improving building energy efficiency.

### **Acknowledgement**

The authors gratefully acknowledge the support of this research by the National Nature Science Foundation of China (Grant No. 71772125), the Natural Science Foundation of SZU (grant no. 2017061) and the Research Grant Council (RGC) of the Hong Kong SAR (152181/14E).

### **References**

- [1] International Energy Agency (IEA), Transition to sustainable buildings: Strategy and opportunities to 2050, July 2013.
- [2] T. Ramesh, R. Prakash, K.K. Shukla, Life cycle energy analysis of buildings: an overview, *Energy and Buildings* 2010; 42: 1592-1600.
- [3] P. Waide, J. Ure, N. Karagianni, G. Smith, B. Bordass, The scope for energy and CO<sub>2</sub> savings in the EU through the use of building automation technology, Final Report for the European Copper Institute, August 2013.

- [4] J.W. Han, M. Kamber, J. Pei, Data mining: Concepts and techniques, 3<sup>rd</sup> edition, 2012, Morgan Kaufman Publishers, MA, USA.
- [5] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Computing Surveys 31 (3) (1999) 264-323.
- [6] E. Rendon, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus external cluster validation indexes, International Journal of Computers and Communications 1 (5) (2011) 27-34.
- [7] A.K. Mann, N. Kaur, Review paper on clustering techniques, Global Journal of Computer Science and Technology Software & Data Engineering 13 (5) (2013)
- [8] L.A. Garcia-Escudero, A. Gordaliza, C. Matran, A. Mayo-Iscar, A review of robust clustering methods, Advances in Data Analysis and Classification 4 (2) (2010) 89-109.
- [9] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks 16 (3) (2005) 645-678.
- [10] A.K. Mann, N. Kaur. Review paper on clustering techniques. Global Journal of Computer Science and Technology Software & Data Engineering 13 (5) (2013) version 1.0.
- [11] S. Vega-Pons, J. Ruiz-Shulcloper, A survey of clustering ensemble algorithms, International Journal of Pattern Recognition and Artificial Intelligence, 25 (3) (2011) 337-372.

- [12] H.P. Kriegel, P. Kroger, A. Zimuk, Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, *ACM Transactions on Knowledge Discovery from Data* 3 (1) (2009) 1-58.
- [13] S. Kotsiantis, D. Kanellopoulos, Association rule mining: A recent overview, *GESTS International Transactions on Computer Science and Engineering* 32 (1) (2006) 71-82.
- [14] J.W. Han, H. Cheng, D. Xin, X.F. Yan, Frequent pattern mining: Current status and future directions, *Data Mining and Knowledge Discovery* 15 (2007) 55-86.
- [15] D. Adhikary, S. Roy, Trends in quantitative association rule mining techniques, In *Proceedings of IEEE International Conference on Recent Trends in Information Systems* (2015) 126-131.
- [16] W. Lian, D.W. Cheung, S. Yiu, An efficient algorithm for finding dense regions for mining quantitative association rules, *Computer & Mathematics with Applications* 50 (3) (2005) 471-490.
- [17] A. Salleb-Aouissi, C. Vrain, C. Nortet, X.R. Kong, V. Rathod, D. Cassard, QuantMiner for mining quantitative association rules, *Journal of Machine Learning Research* 14 (2013) 3153-3157.
- [18] H. Zheng, J. He, G. Huang, Y. Zhang, Optimized fuzzy association rule mining for quantitative data, In *Proceedings of IEEE International Conference on Fuzzy Systems* (2014) 396-403.



- [19] P. Fournier-Viger, C.W. Wu, V.S. Tseng, L. Cao, R. Nkambou, Mining partially-ordered sequential rules common to multiple sequences, *IEEE Transactions on Knowledge and Data Engineering* 27 (8) (2015) 2203-2216.
- [20] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.W. Wu, V.S. Tseng, SPMF: A java open-source pattern mining library, *Journal of Machine Learning Research* 15 (2014) 3389-3393.
- [21] E. Hullermeier, Association rules for expressing gradual dependencies, In *Proceedings of the 6<sup>th</sup> European Conference on Principles of Data Mining and Knowledge Discovery* (2002) 200-211.
- [22] L. Di-Jorio, A. Laurent, M. Teisseire, Mining frequent gradual itemsets from large databases, In *Proceedings of the 8<sup>th</sup> International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII* (2009) 297-308.
- [23] A. Laurent, M.J. Lesot, M. Rifqi, Mining emerging gradual patterns, In *Proceedings of The 16<sup>th</sup> World Congress of the International Fuzzy Systems Association* (2015) 1644-1650.
- [24] T.C. Fu, A review on time series data mining, *Engineering Applications of Artificial Intelligence* 17 (2011) 164-181.
- [25] B. Chiu, E. Keogh, S. Lonardi, Probabilistic discovery of time series motifs, In *Proceedings of the 9<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003) 493-498.

- [26] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: A novel symbolic representation of time series, *Data Mining and Knowledge Discovery* 15 (2007) 107-144.
- [27] D. Minnen, C.L. Isbell, I. Essa, T. Starner, Discovering multivariate motifs using subsequence density estimation and greedy mixture learning, In *Proceedings of the 22<sup>nd</sup> National Conference on Artificial Intelligence* 1 (2007) 615–620.
- [28] Y. Tanaka, K. Iwamoto, K. Uehara, Discovery of time-series motif from multi-dimensional data based on MDL principle, *Machine Learning* 58 (2005) 269–300.
- [29] A. Vahdatpour, N. Amini, M. Sarrafzadeh, Towards unsupervised activity discovery using multi-dimensional motif detection in time series, In *Proceedings of the 21<sup>st</sup> International Joint Conference on Artificial Intelligence* (2009) 1261-1266.
- [30] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gomez-Romero, M.J. Martin-Bautista, Data science for building energy management: A review, *Renewable and Sustainable Energy Reviews* 70 (2017) 598-609.
- [31] M. Goldstein, S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PLoS ONE* 11 (4) (2016) e0152173.
- [32] S. Walfish, A review of statistical outlier methods, *Pharmaceutical Technology* 30 (11) (2006) 82-88.
- [33] H.P. Kriegel, P. Kroger, A. Zimek, Outlier detection techniques, *Tutorials in the 2010 SIAM International Conference on Data Mining*.

- [34] H.P. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data, In Proceedings of the 14<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008) 444-452.
- [35] H.P. Kriegel, P. Kroger, E. Schubert, A. Zimek, Outlier detection in axis-parallel subspaces of high dimensional data, In Proceedings of the 13<sup>th</sup> Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (2009) 831-838.
- [36] A. Lazarevic, V. Kumar, Feature bagging for outlier detection, In Proceedings of the 11<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2005) 444-452.
- [37] A. Zimek, M. Gaudet, R.J.B. Campello, J. Sander, Subsampling for efficient and effective unsupervised outlier detection ensembles, In Proceedings of the 19<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013) 428-436.
- [38] L. Perez-Lombard, J. Ortiz, C. Pout, A review on building energy consumption information, *Energy and Buildings* 40 (2008) 394-398.
- [39] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, *Energy and Buildings* 75 (2014) 109-118.
- [40] X. Ren, D. Yan, T.Z. Hong, Data mining of space heating system performance in affordable housing, *Building and Environment* 89 (2015) 1-13.
- [41] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, P. Patargias, Using principal component and cluster analysis in the heating evaluation of the school building sector, *Applied Energy* 87 (6) (2010) 2079-2086.

- [42] J. Wall, Y. Guo, J.M. Li, S. West, A dynamic machine learning-based technique for automated fault detection in HVAC systems, *ASHRAE Transactions* 117 (2) (2011) 449-456.
- [43] F.W. Yu, K.T. Chan, Assessment of operating performance of chiller systems using cluster analysis, *International Journal of Thermal Sciences* 53 (2012) 148-155.
- [44] C. Fan, F. Xiao, C.C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, *Automation in Construction* 50 (8) (2015) 81-90.
- [45] P. Howard, G. Runger, T.A. Reddy, S. Katipamula, Automated data mining methods for identifying energy efficiency opportunities using whole-building electricity data, *ASHRAE Transactions* 122 (2016) 422-433.
- [46] F.W. Yu, K.T. Chan, Using cluster and multivariate analyses to appraise the operating performance of a chiller system serving an institutional building, *Energy and Buildings* 44 (2012) 104-113.
- [47] G.N. Li, Y.P. Hu, H.X. Chen, H.R. Li, M. Hu, Y.B. Guo, J.Y. Liu, S.B. Sun, M. Sun, Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions, *Applied Energy* 185 (2017) 846-861.
- [48] P.N. Xue, Z.G. Zhou, X.M. Fang, X. Chen, L. Liu, Y.W. Liu, J. Liu, Fault detection and operation optimization in district heating substations based on data mining techniques, *Applied Energy* 205 (2017) 926-940.
- [49] J.J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W.

Tham, K-shaped clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, *Energy and Buildings* 146 (2017) 27-37.

[50] C. Fan, F. Xiao, S.W. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, *Applied Energy* 127 (2014) 1-10.

[51] F. Tang, A. Kusiak, X.P. Wei, Modeling and short-term prediction of HVAC system with a clustering algorithm, *Energy and Buildings* 82 (1) (2014) 310-321.

[52] P.R.S. Jota, V.R.B. Silva, F.G. Jota, Building load management using cluster and statistical analyses, *International Journal of Electrical Power & Energy Systems* 33 (8) (2011) 1498-1505.

[53] D. Hsu, Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data, *Applied Energy* 160 (2015) 153-163.

[54] Z. Yu, F. Haghighat, C.M. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, *Energy and Buildings* 47 (4) (2012) 430-440.

[55] D.F.M. Cabrera, H. Zareipour, Data association mining for identifying lighting energy waste patterns in educational institutes, *Energy and Buildings* 62 (2013) 210-216.

[56] D. Patnaik, M. Marwah, R. K. Sharma, N. Ramakrishnan, Sustainable operation and management of data center chillers using temporal data mining, *KDD* (2009) 1305-1314, Paris, France.

- [57] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, *Automation in Construction* 49 (A) (2015) 1-17.
- [58] C. Miller, Zoltan Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, *Renewable and Sustainable Energy Reviews* (2017), <http://dx.doi.org/10.1016/j.rser.2017.05.124>.
- [59] A. Lavin, D. Klabjan, Clustering time-series energy data from smart meters, *Energy Efficiency* 8 (2015) 681-689.
- [60] F. Iglesias, W. Kastner, Analysis of similarity measures in time series clustering for the discovery of building energy patterns, *Energies* 6 (2013) 579-597.
- [61] V. Figueredo, F. Rodrigues, Z. Vale, J.B. Gouveia, An electricity energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems* 20 (2005) 596-602.
- [62] Z.A. Vale, C. Ramos, S. Ramos, T. Pinto, Data mining applications in power systems case-studies and future trends, *IEEE Transmission & Distribution Conference & Exposition: Asia and Pacific*, 2009, Seoul, South Korea, 2009.
- [63] D. Patnaik, M. Marwah, R. K. Sharma, N. Ramakrishnan, Temporal data mining approaches for sustainable chiller management in data centers, *ACM Transactions on Intelligent Systems and Technology* 2(4) (2011) No. 34.
- [64] H.J. Shao, M. Marwah, N. Ramakrishnan, A temporal motif mining approach to unsupervised energy disaggregation: Applications to residential and commercial buildings, *Association for the Advancement of Artificial Intelligence*, 2013.

- [65] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big BAS data for building energy management, *Energy and Buildings* 109 (2015) 75-89.
- [66] B. Kalluri, A. Kamilaris, S. Kondepudi, H.W. Kua, K.W. Tham, Applicability of using time series subsequences to study office plug load appliances, *Energy and Buildings* 127 (2016) 399-410.
- [67] M. Aiad, P.H. Lee, Unsupervised approach for load disaggregation with devices interactions, *Energy and Buildings* 116 (2016) 96-103.
- [68] M.C. Li, S. Han, J. Shi, An enhanced ISODATA algorithm for recognizing multiple electric appliances from the aggregated power consumption dataset, *Energy and Buildings* 140 (2017) 305-316.
- [69] O. Parson, S. Ghosh, M. Weal, A. Rogers, An unsupervised training method for non-intrusive appliance load monitoring, *Artificial Intelligence* 217 (2014) 1-19.
- [70] M. Aiad, P.H. Lee, Non-intrusive load disaggregation with adaptive estimations of devices main power effects and two-way interactions, *Energy and Buildings* 130 (2016) 131-139
- [71] A.P. Ulmeanu, V.S. Barbu, V. Tanasiev, A. Badea, Hidden Markov Models revealing the household thermal profiling from smart meter data, *Energy and Buildings* 154 (2017) 127-140.
- [72] X.L. Li, C.P. Bowers, T. Schnier, Classification of energy consumption in buildings with outlier detection, *IEEE Transactions on Industrial Electronics* 57 (11) (2010) 3639-3644.
- [73] J.E. Seem, Using intelligent data analysis to detect abnormal energy consumption

in buildings, *Energy and Buildings* 39 (1) (2007) 52-58.

[74] G.J. Lin, D.E. Claridge, A temperature-based approach to detect abnormal building energy consumption, *Energy and Buildings* 93 (2015) 110-118.

[75] V. Jakkula, D. Cook, Outlier detection in smart environment structured power datasets, In the Proceedings of the 6<sup>th</sup> International Conference on Intelligent Environment (2010) 29-33.

[76] I. Khan, A. Capozzoli, S.P. Corgnati, T. Cerquitelli, Fault detection analysis of building energy consumption using data mining techniques, *Energy Procedia* 42 (2013) 557-566.

[77] A. Capozzoli, F. Lauro, I. Khan, Fault detection analysis using data mining techniques for a cluster of smart office buildings, *Expert Systems with Applications* 42 (2015) 4324-4338.

[78] S. Jalori, T.A. Reddy, A new clustering method to identify outliers and diurnal schedules from building energy interval data, *ASHRAE Transactions* 121 (2015) 33-44.

[79] S. Jalori, T.A. Reddy, A unified inverse modeling framework for whole-building energy interval data: Daily and hourly baseline modeling and short-term load forecasting, *ASHRAE* 121 (2015) 156-169.

[80] D.B. Araya, K. Grolinger, H.F. ElYamany, M.A.M. Capretz, G. Bitsuamlak, An ensemble learning framework for anomaly detection in building energy consumption, *Energy and Buildings* 144 (2017) 191-206.

[81] M. Pena, F. Biscarri, J.I. Guerrero, I. Monedero, C. Leon, Rule-based system to



detect energy efficiency anomalies in smart buildings, a data mining approach, *Expert Systems With Applications* 56 (2016) 242-255.

[82] M.D. Jia, R.S. Srinivasan, A.A. Raheem, From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency, *Renewable and Sustainable Energy Reviews* 68 (2017) 525-540.

[83] Z. Yu, C.M. Fung, F. Haghighat, H. Yoshino, E. Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, *Energy and Buildings* 43 (2011) 1409-1417.

[84] A. Capozzoli, M.S. Piscitelli, A. Gorrino, I. Ballarini, V. Corrado, Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings, *Sustainable Cities and Society* 35 (2017) 191-208.

[85] Z. Yu, J. Li, H.Q. Li, J. Han, G.Q. Zhang, A novel methodology for identifying associations and correlations between household appliance behavior in residential buildings, *Energy Procedia* 78 (2015) 591-596.

[86] D. Yan, W. O'Brien, T.Z. Hong, X.H. Feng, H.B. Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: Current state and future challenges, *Energy and Buildings* 107 (2015) 264-278.

[87] J.Y. Yang, M. Santamouris, S.E. Lee, Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings, *Energy and Buildings* 121 (2016) 344-349.

[88] A. Mirakhorli, B. Dong, Occupancy behavior based model predictive control for

building indoor climate – A critical review, *Energy and Buildings* 129 (2016) 499-513.

[89] L.M. Candanedo, V. Feldheim, D. Deramaix, A methodology based on Hidden Markov Models for occupancy detection and a case study in a low energy residential building, *Energy and Buildings* 148 (2017) 327-341.

[90] T.Z. Hong, D. Yan, S. D'Oca, C.F. Chen, Ten questions concerning occupant behavior in buildings: The big picture, *Building and Environment* 114 (2017) 518-530.

[91] S. D'Oca, T.Z. Hong, A data mining approach to discover patterns of window opening and closing behavior in offices, *Building and Environment* 82 (2014) 726-739.

[92] Y. Wang, L. Shao, Understanding occupancy pattern and improving building energy efficiency through Wi-Fi based indoor positioning, *Building and Environment* 114 (2017) 106-117.

[93] L.B. Holder, D.J. Cook, Graph-based data mining, *IEEE Intelligent Systems* 15 (2) (2000) 32-41.

[94] M. Radovanovic, M. Ivanovic, Text mining: Approaches and applications, *Novi Sad Journal of Mathematics* 38 (3) (2008) 227-234.

[95] J. Sivic, A. Zisserman, Video data mining using configurations of viewpoint invariant regions, In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004, Washington, D.C., USA.

[96] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, Towards a sensor for

detecting human presence and characterizing activity, *Energy and Buildings* 43 (2-3) (2011) 305-314.

[97] H.C. Shih, A robust occupancy detection and tracking algorithm for the automatic monitoring and commissioning of a building, *Energy and Buildings* 77 (2014) 270-280.

[98] J.H. Zou, Q.C. Zhao, W. Yang, F.L. Wang, Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation, *Energy and Buildings* 152 (2017) 385-398.